**MACHINE LEARNING-BASED FRAUD DETECTION IN HEALTHCARE INSURANCE WITH OPTIMIZED STORAGE IN BLOCKCHAIN USING ZERO-KNOWLEDGE PROOF**

Arul Arasu N (Roll No: 20Z306)

Dhanaseelan V (Roll No: 20Z313)

Dinesh Baabu R (Roll No: 20Z315)

Lokajit G (Roll No: 20Z328)

Sudarshan S (Roll No: 20Z350)

Dissertation submitted in partial fulfillment of the requirements for the degree of

**BACHELOR OF ENGINEERING**

**Branch: COMPUTER SCIENCE AND ENGINEERING**



APRIL 2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

PSG COLLEGE OF TECHNOLOGY

(Autonomous Institution)

COIMBATORE – 641 004

##### **PSG COLLEGE OF TECHNOLOGY**

(Autonomous Institution)

**COIMBATORE – 641 004**

**MACHINE LEARNING-BASED FRAUD DETECTION IN HEALTHCARE INSURANCE WITH OPTIMIZED STORAGE IN BLOCKCHAIN USING ZERO-KNOWLEDGE PROOF**

Bona fide record of work done by

**Arul Arasu N** (20Z306)

**Dhanaseelan V**  (20Z313)

**Dinesh Baabu R** (20Z315)

**Lokajit G** (20Z328)

**Sudarshan S** (20Z350)

Dissertation submitted in partial fulfillment of the requirements for the degree of

**BACHELOR OF ENGINEERING**

**Branch: COMPUTER SCIENCE AND ENGINEERING**

of PSG College of Technology

**APRIL 2024**

...……………………… …………..……………….  **Dr. V. Santhi**  **Dr.G. Sudha Sadhasivam**

Faculty Guide Head of the Department

Certified that the candidate was examined in the viva-voce examination held on ………………….

…………………….. …………………………..

(Internal Examiner) (External Examiner)

## CERTIFICATE

Certified that this report titled “**MACHINE LEARNING-BASED FRAUD DETECTION IN HEALTHCARE INSURANCE WITH OPTIMIZED STORAGE IN BLOCKCHAIN USING ZERO-KNOWLEDGE PROOF**”, for the Project work II (19Z820) is a bonafide work of

##### **Arul Arasu N (20Z306)**

##### **Dhanaseelan V (20Z313)**

##### **Dinesh Baabu R (20Z315)**

##### **Lokajit G (20Z328)**

##### **Sudarshan S (20Z350)**

who have carried out the work under my supervision for the partial fulfillment of the requirements for the award of the degree of Bachelor of Engineering in Computer Science and Engineering. Certified further that to the best of my knowledge and belief, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion.

**Place: Coimbatore Dr.Santhi V**

**Date: Designation: Professor**

**Department of Computer Science and Engineering**

**PSG College of Technology**

**Coimbatore - 641004**

**COUNTERSIGNED**

**HoD**

**Department of Computer Science and Engineering**

**PSG College Of Technology - 641004**

**ACKNOWLEDGEMENT**

We would like to thank the management of **PSG College of Technology** for providing us the infrastructure and helping us envision new ideas and make them work successfully. We would also like to express our heartfelt gratitude to our **Principal Dr.K.Prakasan** for bestowing us with this valuable opportunity to work in our area of interest.

We also extend our sincere thanks to our **Dr.G.Sudha Sadasivam, Head and Professor, Department of Computer Science and Engineering** for constantly supporting us to develop our ideas and present our project to the faculty committee as a part of our partial fulfillment of the requirements leading to the awarding BE degree.

We take immense pleasure in thanking our project guide **Dr. V. Santhi, Professor and Program Coordinator (G2), Department of Computer Science and Engineering** for being a pillar of support, without whose guidance, unparalleled cooperation and constructive criticisms, this project wouldn’t have been fruitful.

We would also like to thank our tutor **Dr. Vijayalakshmi S, Assistant Professor (Sl. Gr.), Department of Computer Science and Engineering**, for their encouragement and constant vigilance during the course of this project.

A sincere thanks to all the **Panel Members** for reviewing our project and all the **Faculty Members and Staffs** of the department for offering us the required support during the course of the project.

**SYNOPSIS**

The prevalence of fraudulent activities in the insurance industry is alarmingly increasing and requires innovative solutions. The primary objective of the project is to identify instances of fraudulent insurance claims through the utilization of machine learning algorithms and store the claim record in a blockchain network using zero knowledge proofs. Machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest and eXtreme Gradient Boosting are trained with an extensive dataset which is subjected to a sequence of pre-processing techniques and sampling techniques. The trained models are evaluated and compared in terms of different performance metrics. The XGBoost model with SMOTE oversampling is proven to be distinguished among others. The model interpretability is demonstrated by the Explainable AI concept. Blockchain technology is used to store and retrieve the insurance data, ensuring the reliability and immutability of said data. To protect the data from public visibility, zero knowledge proof technology is utilized. This ensures the privacy and authenticity of the information. Overall, this project provides an efficient and automated claim validation system with higher accuracy and efficient storage in blockchain while preventing information leak.

# 

**CHAPTER Page No.**

**ACKNOWLEDGEMENT……..…………………………………………………………3**

**SYNOPSIS……………………………………………………………………………….4**

1. **INTRODUCTION****6**
2. **LITERATURE SURVEY****7**
3. [**SYSTEM REQUIREMENTS**](#_do2khhlbry2f) **17**

3[.1 Software Requirements](#_xzcxmu2vh8uz) 17

3[.2 Hardware Requirements](#_62effme1l5tr) 17

1. **SYSTEM ARCHITECTURE****18**
2. **DATASET DETAILS****20**
3. **IMPLEMENTATION****21**
4. **RESULT AND DISCUSSION…………………………………………………………30**
5. **CONCLUSION…………………………………………………****33**
6. [**BIBLIOGRAPHY**](#_p34d93ai7wmo) **34**

# 

# 

# CHAPTER 1

# INTRODUCTION

In the era of digital world most of the people play some handsome tricks that plays the role the society, organization and the other some sustainable services that could illegally may generate fraud data on foster research of health care services, insurance policy services, bank services and some sustain factors of services that may hazardously incur the fraudulent activities that make economic growth anonymously. These factors affecting the other intellect services of a patient’s information may implant the technology of Machine Learning and Blockchain technology.

Machine learning can be used to detect fraudulent datasets of patient’s insurance claims admitted in a Health care or Hospital that analyzes the behavior patterns and identifies data that need to predict its statistics occurring in the dataset. This could ensure a secure phase from unwanted data that purifies the data to validate on the health care or hospital and insurance policy as well. Blockchain is a highly secure immutable ledger which exists over a distributed network in such a way that only the participants of the network can access the ledger. Provided that the ledger is immutable, no changes can be made to the data added to the blockchain.

This project work collects details about the insurance claim records and moves it to a Node.js environment which acts as an integration between the pretrained machine learning model which makes use of Extreme Gradient Boosting (XGBoost) and Hyperledger Fabric as the blockchain. The insurance claim data is used to predict fraudulent claims which will then be stored in the immutable ledger to prevent such claims in the future. The details of the project work would be discussed in the forthcoming chapters.

Machine learning stands as a fundamental tool in bolstering defenses against insurance fraud, providing a sophisticated system capable of adapting to ever-changing fraudulent strategies which is further secured by the use of Hyperledger Fabric which is an open-source, permissioned ledger framework which acts as a foundation for developing blockchain utilizing applications with a modular architecture. Furthermore, it allows interchangeable components for services like consensus mechanism and membership services in the form of a plug and play environment.

# CHAPTER 2

# LITERATURE SURVEY

Zhou et al in this paper explore the evolving field of blockchain technology in the context of healthcare, with a specific focus on medical insurance. [1] Acknowledging the exponential growth of decentralized and tamper-resistant blockchain systems, the study highlights existing applications in healthcare, such as data-sharing apps and decentralized record management systems. However, it notes a gap in these systems, particularly the absence of homomorphic computing for data recorded on the blockchain. Addressing challenges in the medical insurance business, the paper introduces MIStore, a novel blockchain-based medical insurance storage system. MIStore boasts decentralization, secure data storage using Practical Byzantine Fault-tolerance, a threshold protocol ensuring confidentiality, and verifiability of key data stored in the blockchain. The system also emphasizes efficient verification processes and homomorphic computations to reduce computational overhead. The literature review briefly touches on various blockchain consensus mechanisms and the increasing adoption of blockchain in healthcare. Finally, the paper promises a performance evaluation of MIStore by deploying it on the Ethereum blockchain, showcasing a comprehensive exploration of blockchain's role in revolutionizing medical insurance data management.

Anokye Acheampong Amponsah et al discuss a comprehensive approach to combating healthcare fraud through the integration of machine learning techniques and blockchain technology. [2] Healthcare fraud is identified as a pervasive global issue impacting both developed and developing nations, causing financial losses and hindering the effectiveness of health insurance systems. The proposed solution involves leveraging decision tree classification algorithms to analyze a health insurance claims dataset. The extracted knowledge from these algorithms is then encoded into Ethereum blockchain smart contracts to enhance fraud detection and prevention capabilities. The study emphasizes the applicability of blockchain beyond cryptocurrency, particularly in health insurance, for secure and transparent data management. The research is contextualized within the framework of the National Health Insurance Scheme in Ghana, where financial sustainability threats, including fraud, impede universal health coverage. The integration of machine learning within blockchain smart contracts aims to address this issue. The authors highlight the importance of extending blockchain technology to the insurance sector, specifically in claims processing, showcasing a paradigm shift toward Blockchain 3.0. Experimental results demonstrate the effectiveness of decision tree models, with accuracy reaching 97.96%, sensitivity at 98.09%, and specificity reflecting the system's ability to correctly identify non-fraudulent cases. The study concludes that the proposed system presents a robust solution to detect and prevent healthcare fraud, contributing to the broader goal of achieving Universal Health Coverage.

Shah, D. et al presents a paper which explains a system that integrates machine learning (ML) and blockchain technology. [3] The ML phase involves collecting a dataset from 1540 students and using various classifiers to predict job roles based on academic records. Support Vector Machine and Extreme Gradient Boosting models yielded the best results. The blockchain component utilizes Python classes to represent blocks and utilizes proof-of-work algorithms for data validation. The authors also establish decentralism in the system by creating peer-to-peer networking functionality. Additionally, the paper discusses the integration of blockchain and ML, showcasing an application programming interface that serves requests for predictions. The system aims to provide a new perspective on security and accuracy by combining the verified data available from the blockchain with ML for model training.

B. K. Sethi et al discuss the transformation in the mode of insurance policies from paid per-service to single disease payment. However, there is a possibility of fraud with single-disease payments. [4] In this work, the authors have presented a methodology for detecting the health insurance fraud using blockchain and Machine learning techniques like Support Vector Machine (SVM) and logistic Regression, that can automatically recognize apprehensive medical records to assure sustainable execution of single-disease payment and reduce medical insurance worker's workload. The authors have also proposed a medical record storage and management procedure based on consortium block chain to ensure data security, immutability, traceability, and audit ability. The suggested system may effectively identify fraud and considerably increase the efficiency of medical insurance evaluations, as demonstrated by experiments on two real datasets from two 3A hospitals.

Goyal et al present a blockchain and machine-learning-based framework to revolutionize the health insurance industry, addressing its current challenges of high costs and lengthy claim settlement processes. [5] By utilizing smart contracts and machine learning models like ridge regression and random forests, the proposed system aims to streamline operations, personalize premiums based on individual risk profiles, and significantly reduce the involvement of intermediaries. This will lead to more efficient claim processing, lower premiums, and risk-based premium rebates for policy holders, ultimately enhancing the accessibility and affordability of health insurance. The integration of Internet of Medical Things (IoMT)-based data with blockchain technology furthers the potential for improved data aggregation and decision-making, ensuring patient privacy and system efficiency.

Abbas et al in this paper present a Blockchain and Machine Learning based Framework for Efficient Health Insurance Management which filters the drug supply deploying in the blockchain using Hyperledger Fabric. [6] The pharmaceutical industry faces a significant challenge in combating counterfeit drugs infiltrating the supply chain, resulting in substantial financial losses and potential harm to patients. This paper proposes and implements a novel solution, the Drug Supply Chain Management and Recommendation System (DSCMR), leveraging blockchain technology and machine learning. The blockchain module, deployed using Hyperledger Fabric, ensures efficient monitoring and continuous tracking of drug delivery processes within the smart pharmaceutical industry. Concurrently, the machine learning module utilizes N-gram and LightGBM models trained on a publicly available drug reviews dataset to recommend top-rated medicines to consumers. Integration is achieved through a REST API, enhancing the overall effectiveness and usability of the proposed system. Extensive testing validates the efficiency and reliability of DSCMR in addressing the pressing issue of counterfeit drugs in the pharmaceutical supply chain.

Qiu Zet al emphasize on the privacy and authenticity of the owner’s information in car insurance claims.[7] However, they also point out that the current traditional car insurance claims scenario suffers from inefficiency, complex service, unreliable data, and data leakage. Therefore, they suggest using blockchain, smart contracts, and zero-knowledge proof technology to improve the current problems. This paper proposes a novel car insurance claim scheme based on smart contracts, blockchain, and zero-knowledge proof. It focuses on preserving privacy in the car insurance authorization and claim process. They designed a private smart contract for the creation and revocation of car insurance and a public smart contract for the authorization and validation of car insurance. By using ZoKrates, generating zero-knowledge proofs off chain and verifying the proofs on chain reduces the amount of data storage and computation on chain and provides privacy protection for sensitive information. Experimental results confirm the efficacy of the proposed scheme in terms of security and performance.

S. Jain et al this paper talks about [8] A dual Blockchain structure is employed in this system, utilizing Hyperledger Fabric for granting access to sensitive health data and Ethereum for executing applications and services. The closed nature of Hyperledger Fabric ensures necessary privacy for medical information. The permissioned Blockchain model is highlighted, emphasizing its utility in concealing specific transactions or offers from the public. The system employs a double encryption mechanism for enhanced security, surpassing centralized security systems. The health data security mechanism involves a patient-authorized doctor interaction, where updates to patient records are automatically logged. Access is strictly authorized, and doctors are not permanently granted access; access termination is controlled by the patient. In emergency scenarios, such as unconscious patients, access to health records becomes crucial for informed medical decisions during life saving surgeries. The integration of machine learning is discussed in two steps: dataset adjustment to optimize model accuracy and independent dataset testing for validation and prevention of overfitting. Steps for supervised learning are outlined, including the selection of training set categories, collection of input data and complementary outputs, representation of the input dataset as a feature vector, algorithm selection, and assessment of algorithm accuracy using a test dataset. The overall accuracy of the trained function is dependent on appropriate dataset representation, and the output of the test dataset determines algorithm accuracy.

Ashfaq T et al in this paper address the issues of fraud and anomalies within the Bitcoin network, which are common concerns in e-banking and online transactions. [9] As the financial sector evolves, the methods of committing fraud and anomalies also advance. Despite the introduction of blockchain technology as a secure integration into finance, the number of fraud cases continues to rise each year. To combat this, our proposal suggests a secure fraud detection model that combines machine learning and blockchain. We employ two machine learning algorithms, namely XGboost and random forest (RF), for transaction classification. These techniques train the dataset by analyzing fraudulent and legitimate transaction patterns, enabling them to predict new incoming transactions. By integrating blockchain technology with machine learning algorithms, we can identify and flag fraudulent transactions within the Bitcoin network. In our proposed model, XGboost and random forest (RF) algorithms are utilized for transaction classification and pattern prediction. Furthermore, we evaluate the precision and AUC of these models to measure their accuracy. We also perform a security analysis of the suggested smart contract to demonstrate the robustness of our system. Additionally, we propose an attacker model to safeguard our system against potential attacks and vulnerabilities.

Kayikci et al in this article discuss the integration of blockchain and machine learning technologies. [10] It provides overviews of blockchain concepts like blocks, cryptography, consensus mechanisms, Ethereum, and smart contracts. It also covers machine learning types and processes. The literature review examines studies combining these technologies in IoT, supply chain, medicine, finance, and security. Several real-world examples are described. Some challenges include data standardization, strategic planning, privacy concerns during model training, and scalability. Overall, the combination of these technologies could revolutionize industries by enhancing security, efficiency, and data-driven decision making. However, more work is still needed to address technical and regulatory issues.

Inayatulloh et al aim to make a blockchain application concept of insurance to customer safety risk aided platforms that would be enrolled in the blockchain network. [11] The main framework is provided to web-application for the customer that would be handy to all other risks in future which will be signified to store in their ledger .The insurance participant and customer have been made smart contracts to be the membership of the insurance claim network. It will be handy on any devices such as Computers and Smartphones to access it .

Agrawal D et al discuss using blockchain and machine learning together to decrease the computational time required for training autonomous cars. [12] It proposes training a single "teacher" car with high precision and accuracy, and having that car share its data and continuously updated weights with all other cars through a blockchain network. Further, in the cases of failure or accidents, data is shared among other cars in a secured manner. Also, his integrated approach could reduce the training time needed from being proportional to the number of cars (N) to just 1. It provides examples of how the machine learning and blockchain aspects might be implemented and discusses future directions and challenges.

Nouhaila El AkramI et al in this paper pursue the insight and emerging trends of bibliometric analysis for Web of science core data collection of 700 manuscripts drawn which was spanned from 2017 to 2022. [13] The immense of this technologies that were compiled from various aspects of the research area regarding publication productivity, influential articles, prolific authors, the productivity of academic countries and institutions that are termed in the intellectual structure of using the blockchain and machine learning.The analysis was conducted to provide a valuable foundation for both academic scholars and practitioners using the bibliometric tools to investigate the key area of hotspots, potential prospects, and dynamical aspects of the field.

Sudeep Tanwar et al in this paper imply the great immense technology of Blockchain and Machine Learning that would assist the several smart applications such as Unmanned Aerial Vehicle (UAV), Smart Grid (SG), healthcare, and smart cities. [14] This raised common issues on many directing industry faces on security issues such as majority attack and double-spending. Data analytics is required on blockchain based secure data such that Machine Learning would be analyzed using tradition ML technique such as ,Support Vector Machines (SVM), clustering, bagging, and Deep Learning (DL) algorithms such as Convolutional Neural Network (CNN) and Long short-term memory (LSTM) can be used to analyze the attacks on a blockchain-based network.

Oluwadare Joshua Oyebode et al in this paper discuss the potential of blockchain and machine learning in improving patient care, data management, and clinical research. [15] It highlights the use of blockchain technology in ensuring data reliability and security, as well as its potential to streamline healthcare transactions and reduce costs. The section also emphasizes the role of machine learning in healthcare information management, including the use of predictive analytics to identify high-risk patients and improve treatment outcomes. Additionally, the PDF discusses the potential of the Internet of Things (IoT) in patient monitoring and the use of telemedicine to improve access to healthcare services. Overall, the section showcases the potential for advanced technologies to enhance patient-centric care, drug development, and resource utilization in healthcare settings.

| **Title** | **Technique/**  **Algorithm** | **Metrics** | **Results** | **Scope** |
| --- | --- | --- | --- | --- |
| 1. MIStore:a Blockchain-Based Medical Insurance Storage System | Threshold verifiable homomorphic confidential storage scheme (TVHCSS) | Transaction payload size, Transaction generating time, Throughput | Reduced verification time, for improved throughput with a more suitable blockchain platform | Highlights the areas where blockchain and machine learning are used together, providing a literature overview on the integration of these two technologies, along with their contributions, gaps, and advantages in various fields. |
| 1. A novel fraud detection and prevention method for healthcare claim processing using machine learning and blockchain technology | Decision Tree Algorithm, Ethereum Blockchain | Accuracy, Sensitivity,Specificity, Root mean square error, Mean absolute error , Kappa | Effectiveness of Blockchain-based fraud detection system in making data-driven decisions and significantly reducing the impact of healthcare fraud | Application of Blockchain Technology, Fraud Detection and Prevention,Cost-Efficient Solutions |
| 1. Integrating machine learning and blockchain to develop a system to veto the forgeries and provide efficient results in the education sector. | Multiple Machine Learning models, Blockchain built from scratch | Mean absolute error, Root mean square error, Relative absolute error and Relative squared error | Support Vector Machine (SVM) and XGBoost provides highest accuracies. Successful creation and integration of blockchain with the model. | Use of existing blockchain technologies.  Reduction in Time complexity for data retrieval. |
| 1. Medical Insurance Fraud Detection Based on BlockChain and Machine Learning Approach | Consortium Blockchain, Support Vector Machine, Logistic Regression | Accuracy  Sensitivity  Specificity | * Blockchain and ML for fraud detection in single-disease payments, enhancing efficiency. * Consortium blockchain ensures data security and traceability, improving medical insurance evaluations. | * Explore scalability and interoperability for broader healthcare industry applications. * Evaluate real-world implementation challenges and user acceptance for practicality. |
| 1. A Blockchain and Machine Learning based Framework for Efficient Health Insurance Management | Ridge regression, Random Forest, Blockchain network | R-squared, Root mean square error, Mean absolute error, Quadratic weighted Kappa | * Accurate, cost-effective and quick management of insurance claims. * Use of fewer intermediaries and smart contracts. | Improve the accuracy in prediction of the insurance premium.  Make the scheme lightweight to reduce communication overheads. |
| 1. A Blockchain and Machine Learning-Based Drug Supply Chain Management and Recommendation System for Smart Pharmaceutical Industry. Electronics. | N-gram, LightGBM, Sentiment Analysis , Hyperledger Fabric | Accuracy, Recall, Precision, F1 Score, Execution Time | Drug Supply Chain Management. Drug Recommendation Results,Efficiency and Usability Testing . These results showcase the functionality and potential impact of the proposed system, highlighting its ability to provide drug recommendations and enhance the efficiency | The system provides secure and transparent drug supply chain management while also recommending the accurate medicines. |
| 1. Novel Blockchain and Zero-Knowledge Proof Technology-Driven Car Insurance | * Blockchain * Zero-Knowledge Proof * ZoKrates | Number of Constraints and Key SizeTime CostGas ConsumptionCharacteristic Comparison | Proposal of a hybrid smart contract proxy model  * Utilization of ZoKrates for zero-knowledge authorization and verification | * Scalability of blockchain technology * Efficiency of zero-knowledge proof algorithms * Optimize the performance of the zero-knowledge proof algorithm * Implement the model in an automated manner |
| 1. Blockchain and Machine Learning in Health Care and Management | Hyperledger and ethereum blockchain technologies. | Read latency  Read throughput  Transaction latency  Transaction throughput | Improved security, privacy, and data integrity in managing sensitive health data. | Optimize model accuracy for various healthcare applications |
| 1. A Machine Learning and Blockchain Based Efficient Fraud Detection Mechanism | * Blockchain Technology * Machine Learning Algorithms (XGboost and random forest) * Deep Learning Algorithms (Bidirectional Long Short-Term Memory - BiLSTM) * Encryption Techniques (Asymmetric, symmetric, and homomorphic encryption) | precision-recall curve, accuracy, log loss, and area under the curve (AUC) | Proposal of a Blockchain-based Machine Learning Algorithm  Address Data Imbalance by generating synthetic malicious data points | * The observed log loss of XGBoost during training indicates efficient capture of nonlinear patterns * the precision-recall curve shows optimal accuracy in classifying blockchain transactions. |
| 1. Blockchain meets machine learning | K-nearest neighbor (K-NN) model for cryptocurrency price prediction, Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) | Precision, recall, F-score, root mean square error, convergence rate, data latency, throughput, and transaction latency | Delves into the use of various algorithms and performance metrics in the context of blockchain and machine learning. | It provides insights into the technical aspects of blockchain technology, its potential applications in different domains, and the emerging research in the field of medicine, finance, and security. |
| 1. Blockchain technology of fraud Detection and Risk Prevention in Insurance Industry | Ethereum,Blockchain network. | Transaction speed, scalability, and security | Web Application on devices such as Computers and Smartphones. | Portable, easy to store data without loss, accessiblilty |
| 1. Blockchain Integrated Machine Learning for Training Autonomous Cars | Q-Learning, Attention U-Net framework, and self-writing smart contracts. | Data latency reduction, convergence rate improvement, storage optimization, precision | Integration of blockchain and machine learning technologies, highlighting their potential applications in various domains such as finance, medicine, supply chain, and security | Provides insights into the benefits, challenges, and real-world examples of combining these technologies. |
| 1. Unleashing the Potential of Blockchain and Machine Learning: Insights and EmergingTrends From Bibliometric Analysis | Bibliometric tools for Bibliometric Analysis | Wos - Web of Science (Data collection) | Searching the bibliography data gives experts and scholars details and cite counts. | Connecting scholars and experts in academics for specified fields and domain. |
| 1. Machine Learning Adoption in Blockchain-Based Smart Applications: The Challenges, and a Way Forward | Support Vector Machines (SVM), Deep Learning (DL) algorithms such as Convolutional Neural Network (CNN) and Long short-term memory (LSTM) | Recall, F1 score, root mean square error, accuracy, throughput, and transaction times | Gives security to those issues of cyber-attack in blockchain and analyze the attacks on a blockchain-based network. | Both technologies applied several smart applications such as Unmanned Aerial Vehicle (UAV), Smart Grid (SG), healthcare, and smart cities. |
| 1. Blockchain and Machine Learning Based Healthcare | Artificial neural networks (ANNs) and chemical reaction optimization (CRO) | Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) | Highlights the increasing interest and publications in this domain, indicating the growing application of these technologies in healthcare | Explore the utilization of machine learning and blockchain technologies in the field of smart healthcare. |

# CHAPTER 3

# SYSTEM REQUIREMENTS

## 3.1 Software Requirements

**Blockchain Platform:** A suitable blockchain platform for implementing the system - Hyperledger Fabric.

**Smart Contracts:** To handle the validation of insurance claims and their addition to the blockchain. Further, these contracts should include logic for verifying claim details and authenticity.

**Machine Learning Models:** Integrate machine learning models for fraud detection. These models can analyze historical data to identify patterns of fraudulent claims.

## 3.2 Hardware Requirements

**Server Infrastructure:** Depending on the scale of your system, server infrastructure will be needed to host the blockchain nodes, the DApp backend, and any other supporting services.

**Cloud Services:** Use of cloud platforms like AWS, Azure, or Google Cloud for scalability, reliability, and ease of management.

**Security Measures:** Implement hardware security modules (HSMs) to enhance the security of private keys and cryptographic operations.

**Data Storage:** Set up storage solutions for storing blockchain data, historical data, and backups.

**Networking:** Ensure a reliable and fast internet connection to support the communication between nodes and users

# CHAPTER 4

# SYSTEM ARCHITECTURE

# 

**Fig 4.1 - System Architecture**

In the proposed system architecture Fig 4.1, we focus on seamlessly integrating machine learning and blockchain technologies to enhance insurance fraud detection while prioritizing privacy through the implementation of zero-knowledge proofs. At the core, the machine learning component processes and analyzes insurance claims data, identifying potential fraud patterns. This information is then securely transmitted to the blockchain, utilizing zero-knowledge proofs to validate the accuracy of the detected fraud without revealing sensitive details. The blockchain acts as an immutable ledger, ensuring transparency and traceability of the insurance claims. Smart contracts, embedded within the blockchain, execute fraud detection logic and trigger alerts when fraudulent activities are identified. This architecture not only facilitates a robust fraud detection mechanism but also upholds data privacy by leveraging zero-knowledge proofs, ensuring that only essential information is disclosed while maintaining the integrity of the insurance claims process.

# 

**Fig 4.2 - Work Flow Diagram**

Fig 4.2 shows the work flow diagram where users submit insurance claims, which are securely recorded on the blockchain. Zero-knowledge proofs are employed to validate the authenticity of claims without revealing specific details.The machine learning module processes the claims, identifying potential fraud patterns based on historical data and evolving threat landscapes.The fraud detection smart contract triggers based on the machine learning analysis, executing predefined actions such as flagging suspicious claims or requesting additional verification.Confirmed results, along with necessary proof, are securely recorded on the blockchain, creating an immutable audit trail for each insurance claim.

# CHAPTER 5

# DATASET DETAILS

For the purpose of this project, we have taken a sample dataset from Kaggle. It consists of Inpatient claims, Outpatient claims and Beneficiary details of each provider. This project uses eight data files out of which 4 are for training purposes and 4 are used in testing the model. There are 4 types of data where each type has one file for training and testing. The types of data file are - A) Inpatient Data - This dataset offers valuable information on the insurance claims submitted for individuals who have been admitted to hospitals. It includes additional details such as admission and discharge dates, as well as the admit diagnosis code, B) Outpatient Data - This dataset provides information on insurance claims for patients who visit hospitals but are not admitted. It captures relevant details associated with their hospital visits, C) Beneficiary Details Data - This dataset encompasses KYC details of beneficiaries, including health conditions and the region they are affiliated with, and D) Provider Data - This file contains Provider ID. In the training file, the Provider ID is mapped with target value - Potential Fraud, whereas in the testing file only the Provider ID is given.

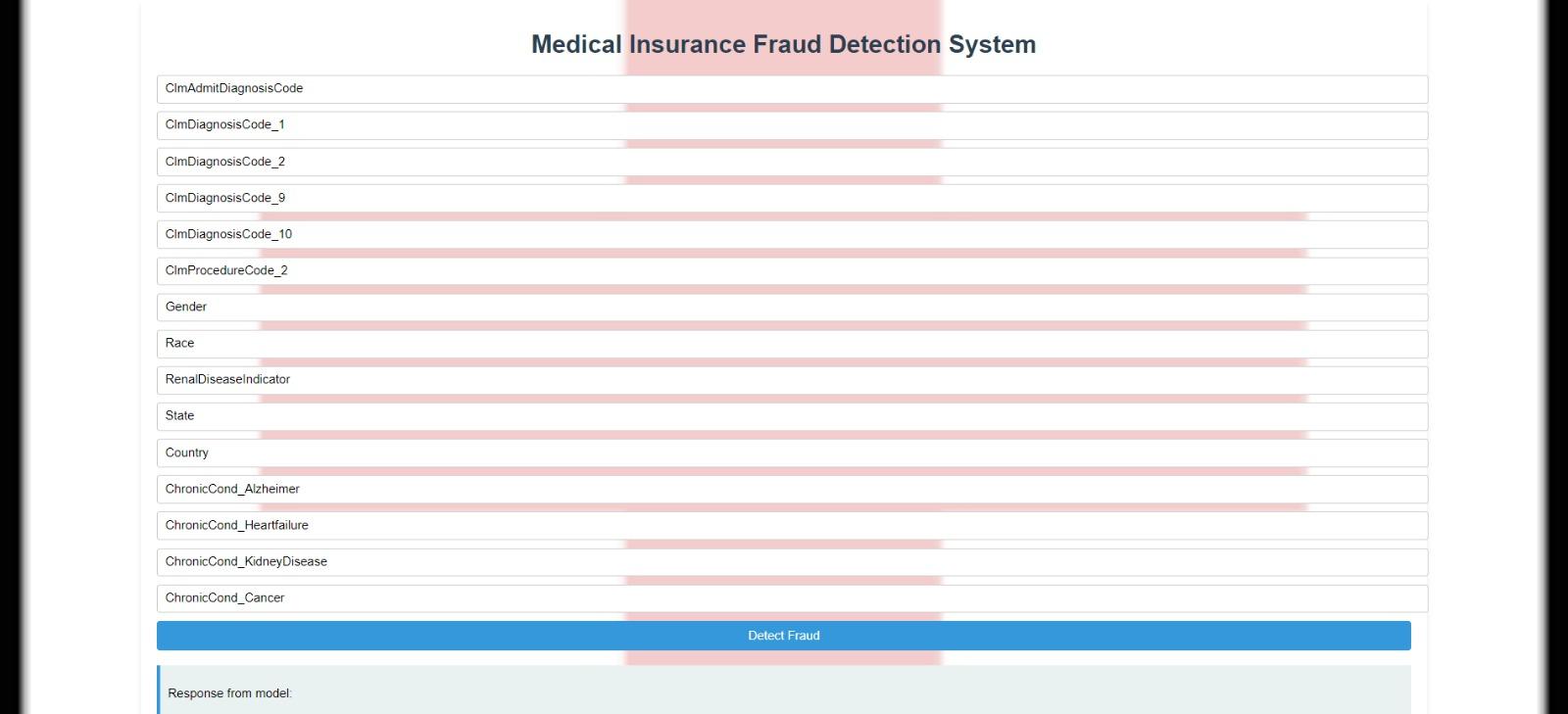
<https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis>

# CHAPTER 6

# IMPLEMENTATION

**6.1 FRONT END UI FOR ACCEPTING INPUT FROM USER**

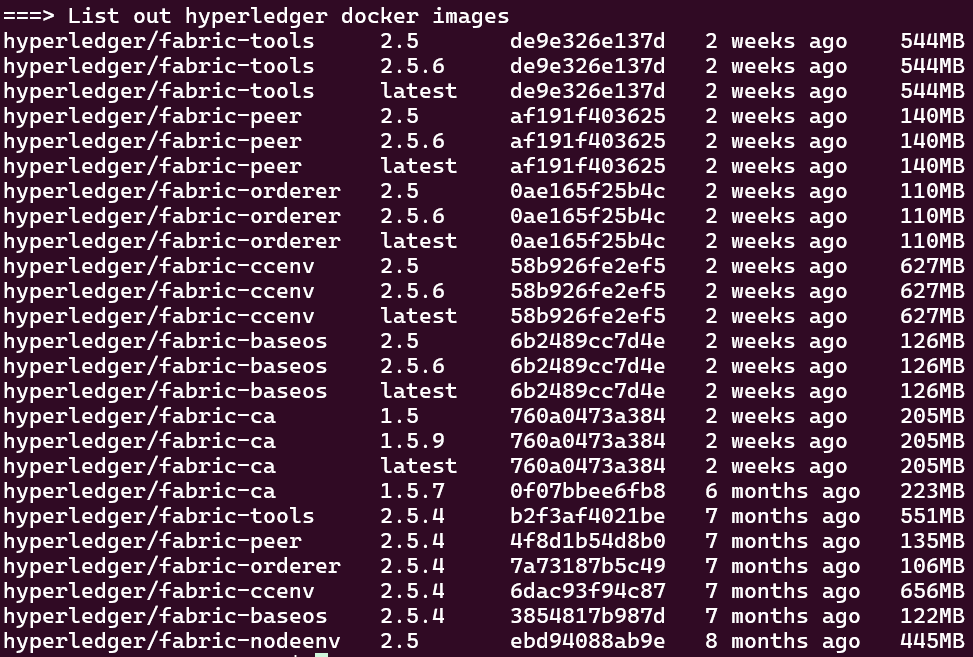
With React, users can easily input the important features. Each feature, ranging from age to gender and medical conditions, is presented with clear labels and input boxes. As users type, the form updates automatically. It's designed to be easy to read and use. Once users finish inputting their information, the data is used for detecting if the record is fraudulent or not. This user interface is shown in Fig 6.1.



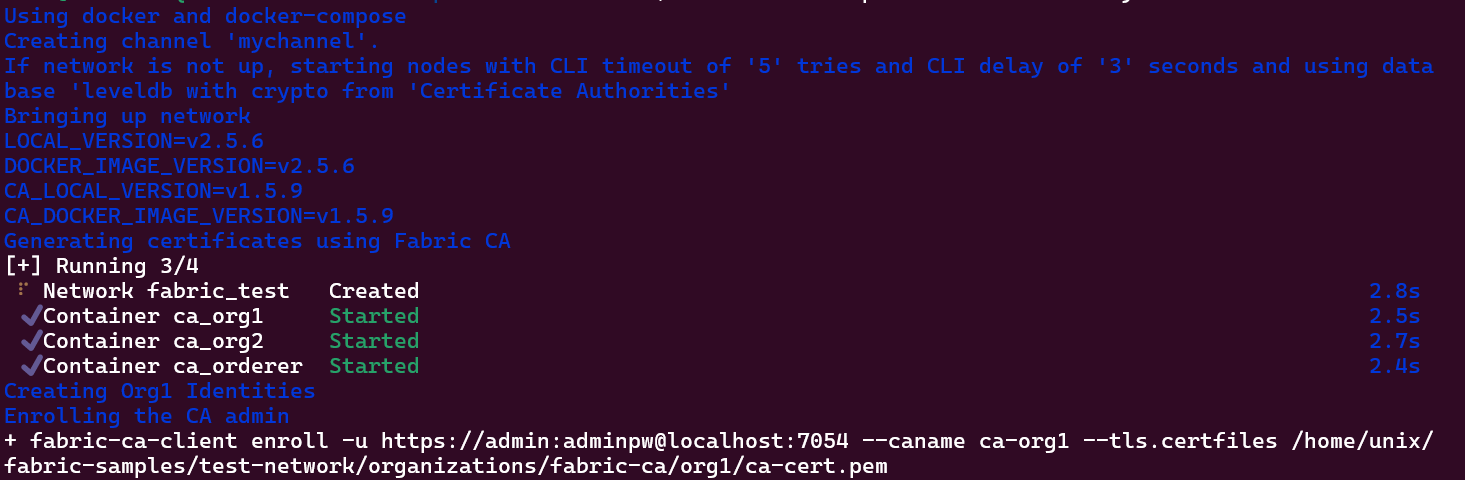
**Fig 6.1 - Form to get user input**

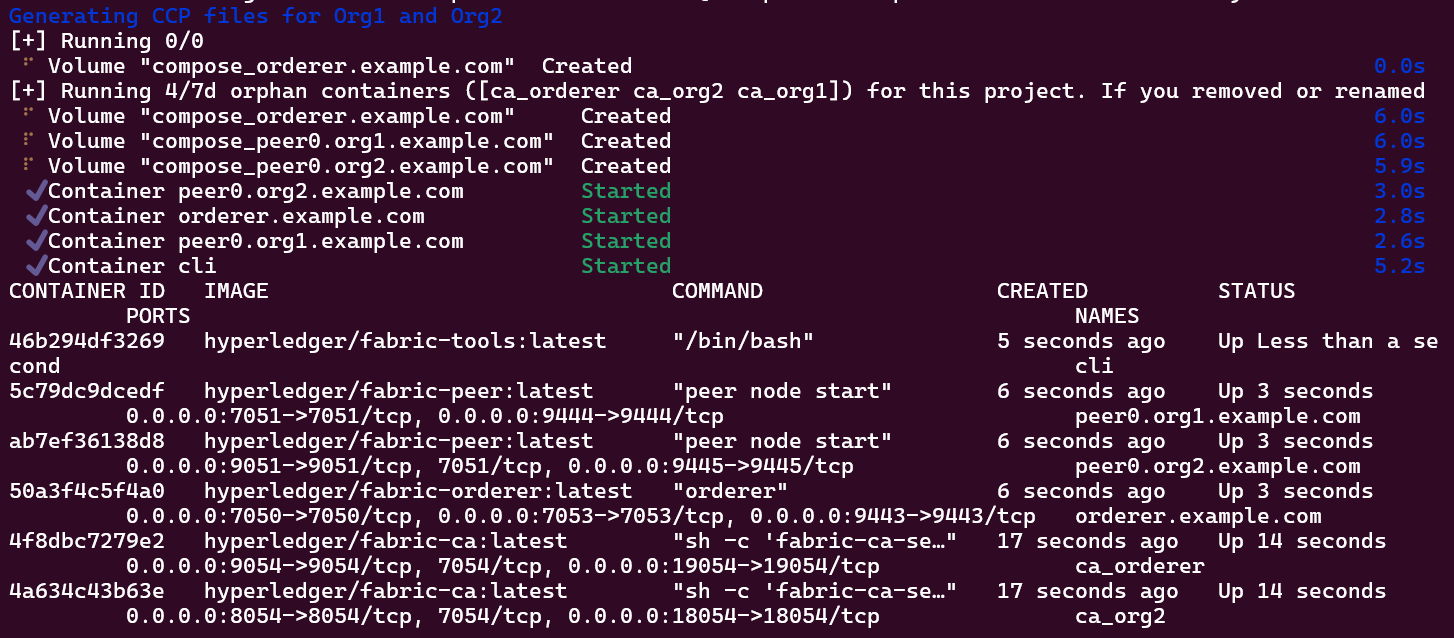
**6.2 PREREQUISITES FOR USING HYPERLEDGER FABRIC**

Channels in Hyperledger Fabric serve as a means to create private and confidential communication pathways between a subset of network members. They enable different parties to transact securely and privately, shielding their transactions and data from the rest of the network participants. Channel creation can be performed parallel to the bringing up of the blockchain network. During this stage, a choice between the use of a cryptogen tool or a certificate authority is also available. Cryptogen is a command-line tool provided for generating cryptographic keys and X.509 certificates, necessary for establishing secure communication channels among network participants and for further authentication. On the other hand, certificate authorities are trusted entities responsible for issuing, revoking, and managing digital certificates in order to verify the identity of network participants. With the use of certificate authorities, the registration of the admin user is also performed internally. Chaincode is the smart contract paradigm of Hyperledger fabric and it represents the business logic or application logic that defines the rules and procedures governing interactions within the blockchain network. Chaincode is instantiated and deployed to the peers within the network, enabling them to execute transactions and update the ledger's state in a deterministic and secure manner. Each chaincode instance operates within the context of a specific channel, ensuring data isolation and confidentiality among network participants. This is used to tokenize the features like State, County, Presence of a chronic condition like Alzheimers, Claim admit diagnosis code corresponding to the data obtained from insurance claims and the prediction results provided by the machine learning model. Then, the chaincode is deployed through the channel and is defined in all the nodes which are part of the channel.

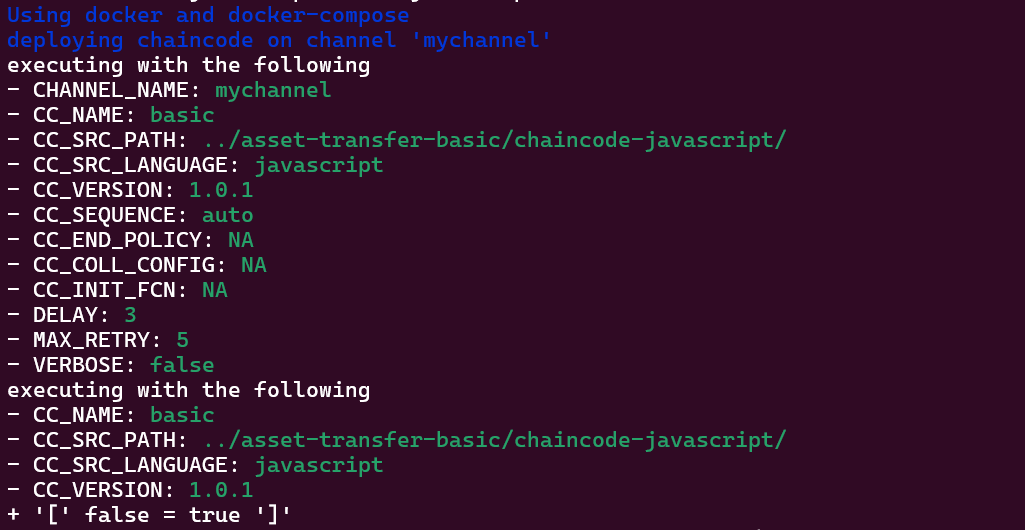


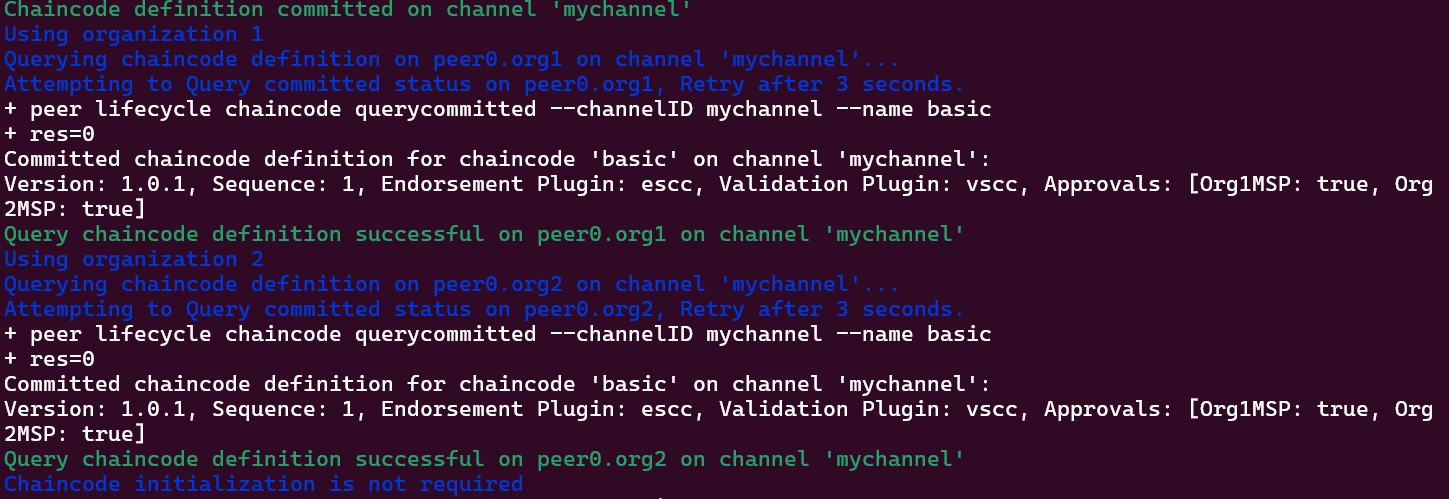
**Fig 6.2 - Hyperledger Fabric Installation**





**Fig 6.3 - Channel creation**

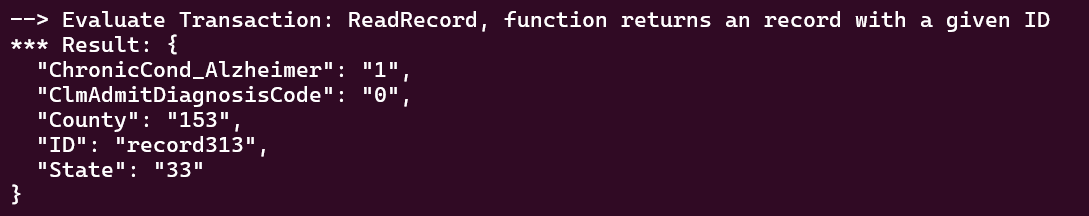




**Fig 6.4 - Chaincode deployment and initialization**

**6.3 USING HYPERLEDGER FABRIC THROUGH NODE.JS AND EXPRESS**

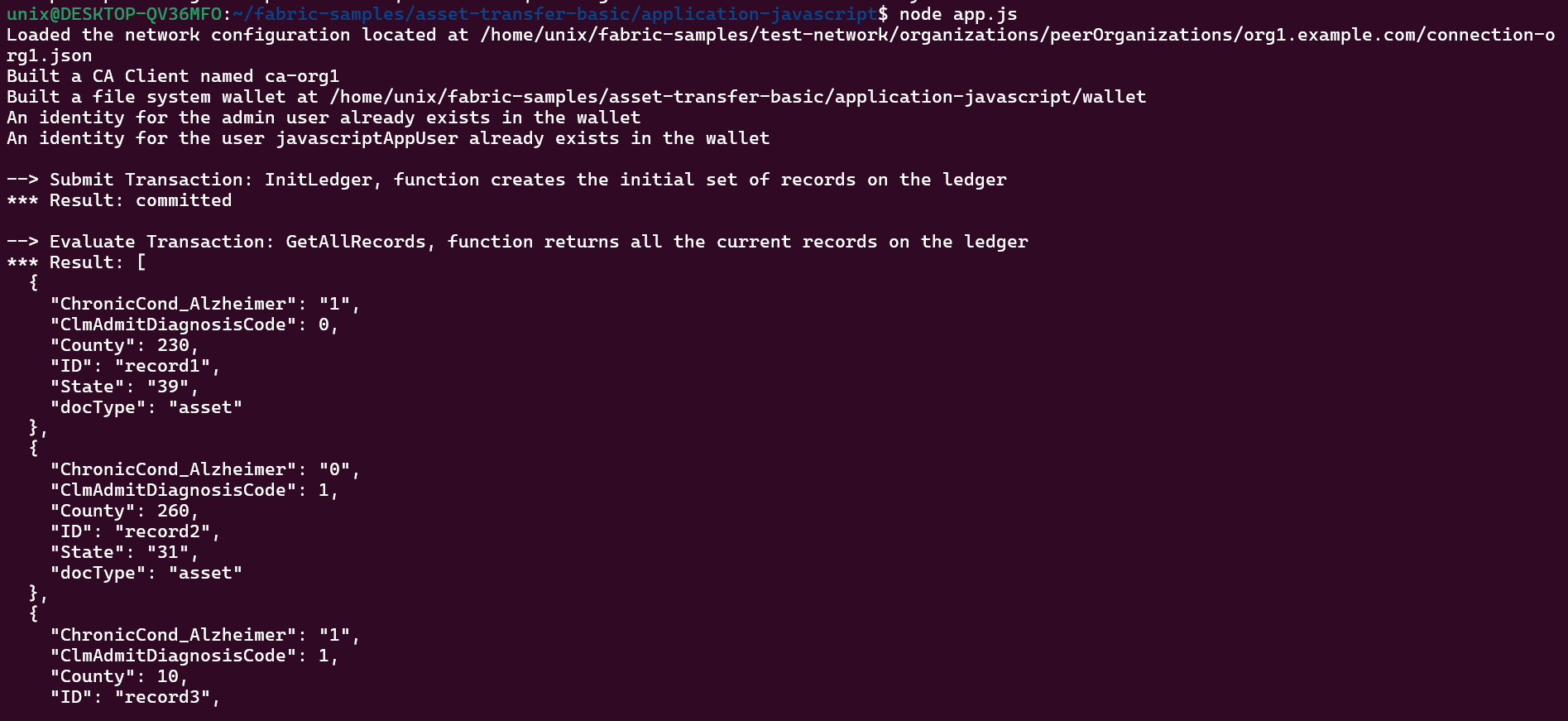
The functions provided by the deployed chaincode are made available to invoke through the Fabric contract API. A simple Node.js application is created and is instantiated with the blockchain network with separate variables for the channel, chaincode, gateway and membership service provider. The gateway is used to communicate with the fabric network. A separate folder for wallet is created with unique identifiers for the admin and the user of the application. Then, an instance for the network is created from the gateway and subsequently an instance of the chaincode is created from the network instance. This makes the hyperledger fabric available for querying using functions like submitTransaction() and evaluateTransaction(). Three different query functions have been created. A getAllAssets() function to get all the records added to the hyperledger fabric network (Fig 6.7), a createAsset() function is used to create new records to be added to the blockchain, a readAsset() to read the items found within a particular block (Fig 6.5) and a assetExists() function to verify the existence of a particular block with its identifier (Fig 6.6). This can be exposed to the user through express (Fig 6.8).



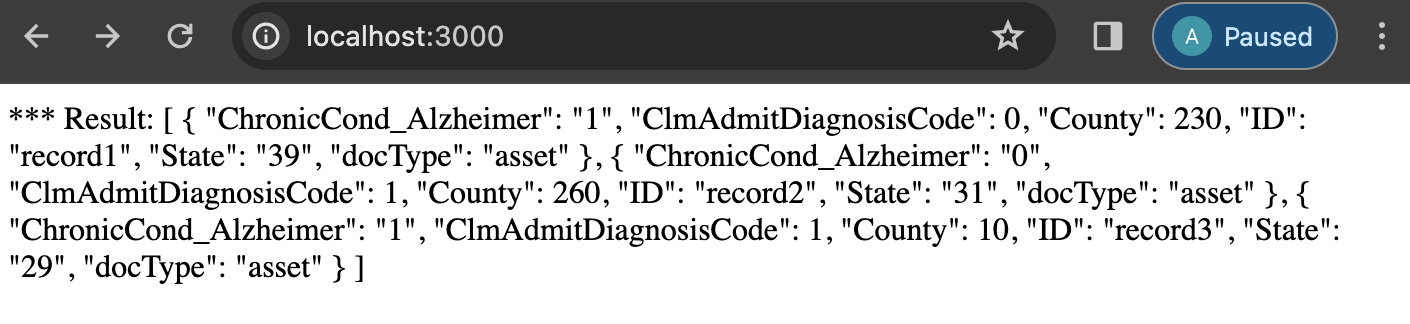
**Fig 6.5 - readAssset()**



**Fig 6.6 - recordExists()**

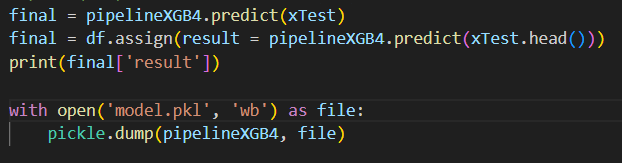
****

**Fig 6.7 - getAllAssets()**

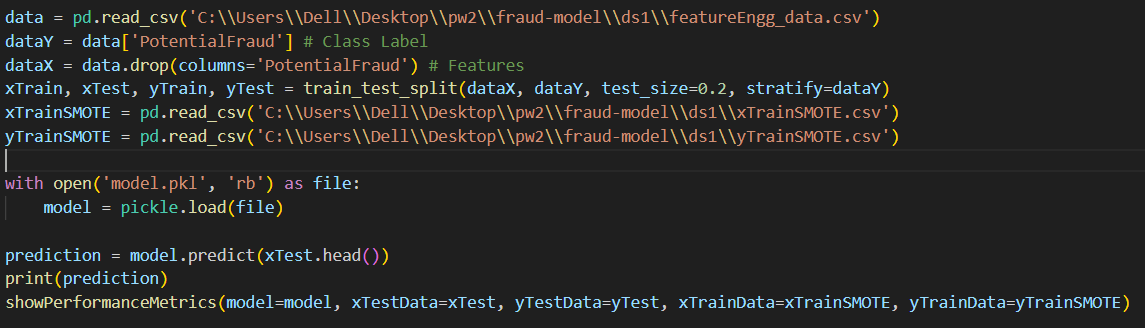
**Fig 6.8 - Exposed through express**

**6.4 PICKLING THE MACHINE LEARNING MODEL**

The pretrained machine learning model which makes use of the Extreme Gradient Boosting Algorithm (XGBoost) for classification of response encoded and standardized data is dumped into a .pkl file using the pickle library (Fig 6.9). The dumped file can be loaded across any python environment. The pickled file which comprises the pretrained machine learning model is imported into a different python script in order to avoid repetitive training of the model in the same data which inevitably results in overfitting. Upon loading of the pickled file, the predictions performed are to be imported into the Node.js environment which projects the communication between the machine learning model and the blockchain network (Fig 6.10). Currently, the output provided by the model is exposed through the child\_process module which is used by the Node.js application interacting with the blockchain.



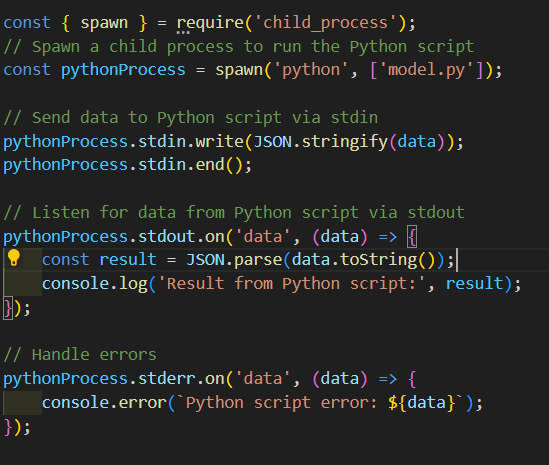
**Fig 6.9 - Saving the model**



**Fig 6.10 - Loading the saved model**

**6.5 EXPOSING PRETRAINED MODEL USING CHILD\_PROCESS MODULE**

The machine learning model is deployed, as in it is made accessible to the Node.js application running on top of the Hyperledger blockchain using the child\_process module. The Child Process module in Node.js allows for the execution of other programs or scripts from within a Node.js application (Fig 6.11). It provides the functionality to spawn subprocesses, which can run asynchronously and communicate back with the parent process via messaging or by using standard input and output streams. This module is crucial for handling operations that require external system commands or need isolation from the main application flow. This functionality of the child process module has been utilized in such a way that the blockchain interactions occur as part of the parent process. The python script which imports the pickled machine learning model is called as a child process using the spawn() function call. The data obtained from the user is serialized to become a JSON-formatted string which will be used by the machine learning model to produce an output. The obtained output which is a pandas dataframe is parsed and converted into a string before shown as the result to the user’s query. To observe the working of this particular module, the output is printed on the terminal (Fig 6.12).



**Fig 6.11 - Using child\_process library**



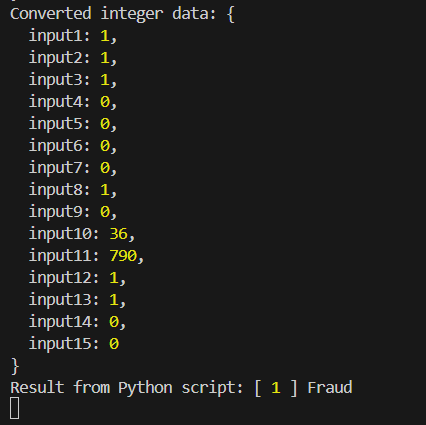
**Fig 6.12 - Output from terminal**

**6.6 FINAL PROCESS FLOW**

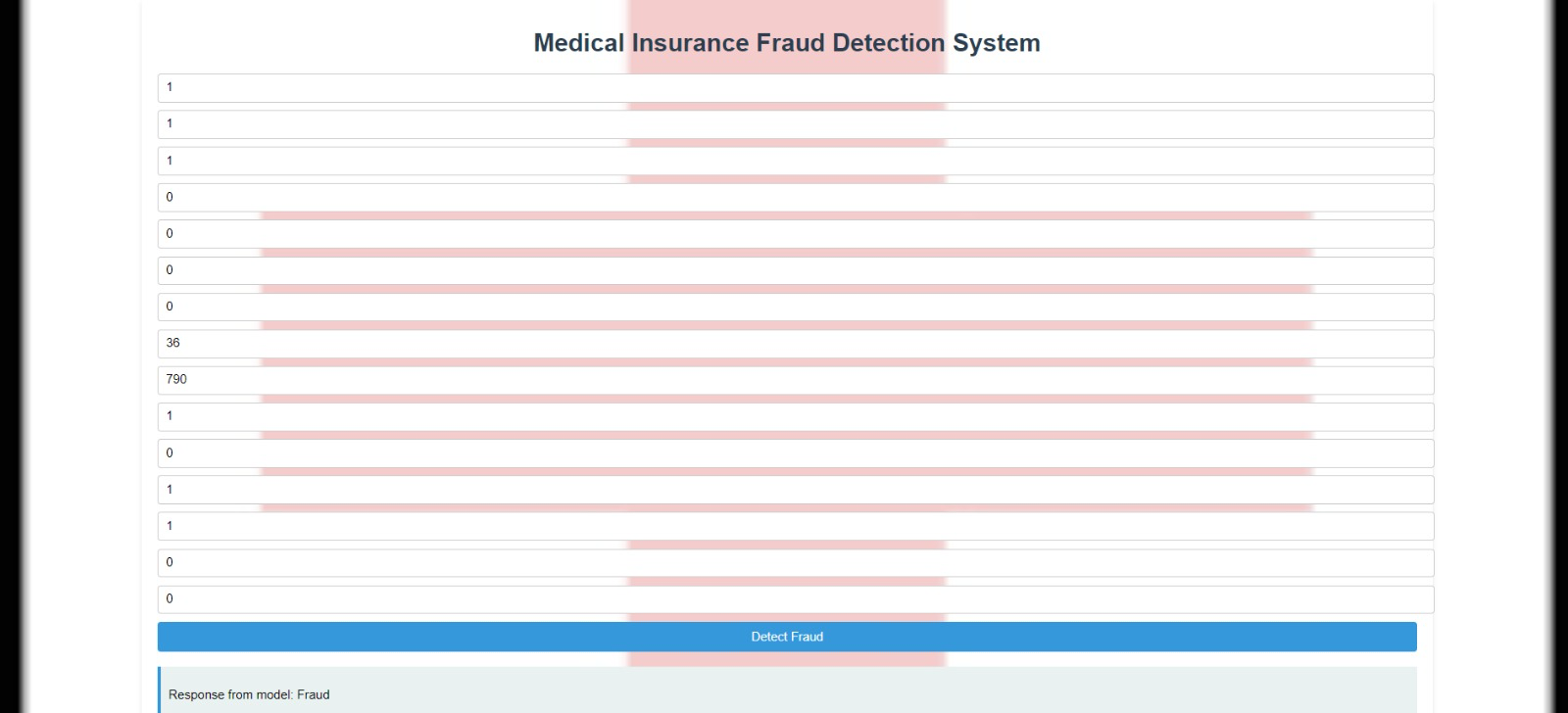
To summarize, as part of the overall process flow, the user provides sensitive information about the insurance claims through the front end UI. This data is then made accessible for the machine learning model through the child\_process module from the Node.js application running on top of the Hyperledger Fabric (Fig 6.13). Based on the predictions made by the model, the classification is made visible to the user while also being added to the fabric blockchain where it becomes a part of an immutable ledger. The input data from the frontend as received in the backend can be seen in Fig 6.14. The final output classification based on the insurance claim data is notified to the user before additions to the blockchain are made (Fig 6.15).



**Fig 6.13 - Sending data to the backend**



**Fig 6.14 - Input seen from the backend**

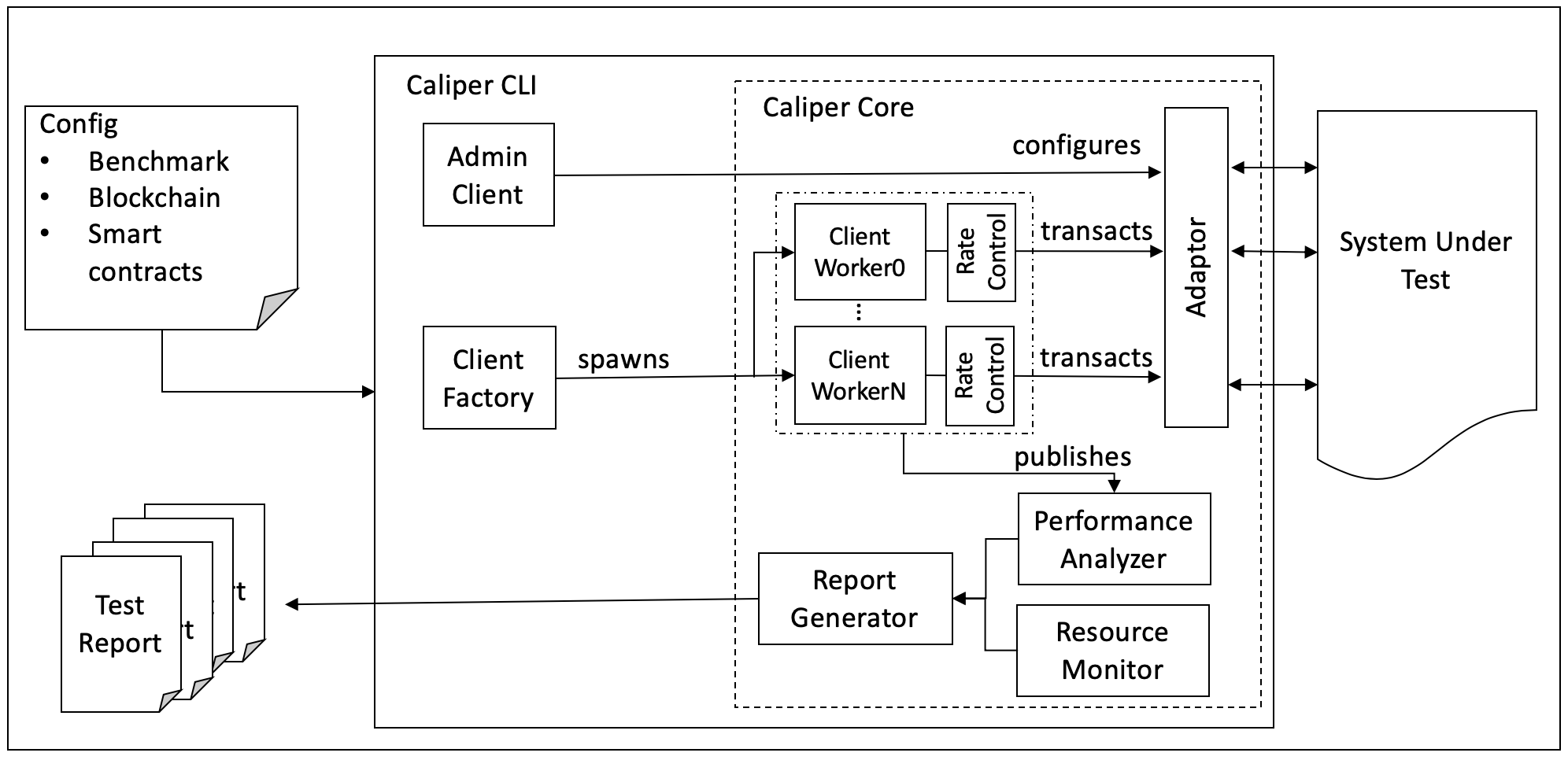


**Fig 6.15 - Final output**

**CHAPTER 7**

# RESULT AND DISCUSSION

Hyperledger Caliper is a powerful blockchain benchmarking tool used to evaluate the performance of blockchain networks. With Caliper, developers can simulate various blockchain scenarios and measure the system's throughput, latency, and resource utilization. It supports multiple blockchain platforms, including Hyperledger Fabric, Sawtooth, and Ethereum, making it versatile for different blockchain implementations. Caliper's modular architecture (Fig 7.1) enables easy customization and extension for specific testing requirements. By providing standardized benchmarks and metrics, Caliper helps developers identify performance bottlenecks and optimize their blockchain applications. Its comprehensive reporting capabilities allow for in-depth analysis and comparison of different blockchain configurations. Caliper plays a crucial role in ensuring the scalability, reliability, and efficiency of blockchain solutions in diverse real-world scenarios.



**Fig 7.1 - Caliper architecture**

**7.1 TESTING WITH HYPERLEDGER CALIPER**

The usage of Caliper has enabled the measurement of critical metrics, such as transaction per second (TPS), send rate, and transaction latency, providing a detailed overview of the blockchain's operational efficiency. By executing defined transactions against their network, precise data on throughput and resource consumption under various network conditions can be captured. The resultant report generated by Caliper not only highlights these metrics but also helps in identifying potential bottlenecks and areas for improvement. This ensures that the blockchain solution is not only robust but also optimized for high performance, offering tangible evidence of its scalability and reliability. Thus, the performance of the blockchain in terms of efficiency and scalability is assessed.

Summary

| Name | Success | Failure | Send rate (TPS) | Max Latency (s) | Min Latency (s) | Avg Latency (s) | Throughput (TPS) |
| --- | --- | --- | --- | --- | --- | --- | --- |
| open | 1000 | 0 | 57.4 | 6.48 | 0.68 | 3.49 | 52.4 |
| open | 1000 | 0 | 60.3 | 7.89 | 1.14 | 4.46 | 49.9 |
| query | 2000 | 0 | 165.8 | 9.32 | 0.07 | 4.81 | 122.7 |
| transfer | 0 | 2000 | 97.4 | 0.00 | 100000.00 | NaN | 0.0 |

**Fig 7.2 - Performance metrics**

Round 0

Performance metrics

| Name | Success | Failure | Send rate (TPS) | Max Latency (s) | Min Latency (s) | Avg Latency (s) | Throughput (TPS) |
| --- | --- | --- | --- | --- | --- | --- | --- |
| open | 1000 | 0 | 57.9 | 6.58 | 0.71 | 3.53 | 52.1 |

Resource consumption

| Type | Name | Memory (max) | Memory (avg) | CPU% (max) | CPU% (avg) | Traffic in | Traffic out | Disc read | Disc write |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Docker | compose-peer0.org2.example.com-v0 | 6.3 MB | 6.3 MB | 0.02 | 0.00 | 706 B | 66 B | 0 B | 0 B |
| Docker | compose-peer0.org1.example.com-v0 | 6.2 MB | 6.3 MB | 0.00 | 0.00 | 706 B | 66 B | 0 B | 0 B |
| Docker | compose-peer0.org2.example.com-simple-v0 | 7.0 MB | 6.7 MB | 5.43 | 2.37 | 1.4 MB | 645 KB | 0 B | 0 B |
| Docker | compose-peer0.org1.example.com-simple-v0 | 7.2 MB | 6.8 MB | 5.93 | 2.48 | 1.4 MB | 652 KB | 0 B | 0 B |

Send rate refers to the number of transactions which can be sent per unit time. Given that the send rate values are high, it means that the application is performing well. Throughput is used to measure the amount of data or transactions that can be processed within a given time period. The throughput for different operations range between 49 to 120. Latency refers to the time taken for a transaction or a data packet to travel from its source to the destination in this case, the transaction moving from its origin point into the blockchain. The average latency is a significantly lower value. Further, the resource utilization of every docker image can be seen in the bottom table which only extends up to the megabyte range. Further, the memory consumption of the individual docker images exhibit only CPU utilization of about 2% is required.

# CHAPTER 8

# CONCLUSION

In conclusion, this project successfully adopts a machine learning model along with blockchain technology for the detection of fake insurance claims, a critical issue in the insurance industry. The machine learning model used in XGBoost coupled with SMOTE Oversampling with a F1 score of 0.76733. This model is paired with Hyperledger Fabric in a Node.js environment. The frontend user interface acts as the direct communication medium between the user and the entire integration between the machine learning model and the blockchain.

After the necessary connections were made, the entire application as a whole was tested using Hyperledger Caliper.It ended up being essential in testing this project, ensuring the performance and reliability. With thorough testing, Caliper provided various performance metrics which gave valuable insights into the system's capabilities and it has also helped in optimizing the solution and ensuring its seamless operation, in terms of metrics like send rate, throughput and latency for different types of transactions.

Upon testing for performance metrics, the send rate of the transactions lies in the range of 50 to 166 per second, the latency varied between 0.07 to 9.27 and throughput varied between 49.9 to 122.4. Observing that the send rate and throughput values are significantly higher while the latency is lower means that the application is working in an efficient manner.

This application will greatly benefit the insurance industry, enhancing its ability to combat fraudulent activities and protect the interests of both insurers and policyholders. Furthermore, this research underscores the importance of selecting the right combination of machine learning algorithms and feature selection techniques to achieve optimal results in complex, real-world problem-solving scenarios, along with a secure storage of obtained information in the blockchain which promotes confidentiality and security.

# CHAPTER 9

# BIBLIOGRAPHY

1. Zhou, L., Wang, L. & Sun, Y. MIStore: a Blockchain-Based Medical Insurance Storage System. J Med Syst 42, 149 (2018). https://doi.org/10.1007/s10916-018-0996-4
2. Anokye Acheampong Amponsah, Adebayo Felix Adekoya, Benjamin Asubam Weyori, A novel fraud detection and prevention method for healthcare claim processing using machine learning and blockchain technology, Decision Analytics Journal, Volume 4, 2022, 100122, ISSN 2772-6622, <https://doi.org/10.1016/j.dajour.2022.100122>.
3. Inayatulloh, Siti Elda Hiererra, Prasetya Cahaya S, Rozil Toyob, Nico Djundharto Djajasinga , Sawqi Saad El Hasan, Rofiq Noorman Haryadi, Rivaldhy N. Muhammad,“ Blockchain technology of fraud Detection and Risk Prevention in Insurance Industry “https://ieomsociety.org/proceedings/2022paraguay/47.pdf”
4. Shah, D., Patel, D., Adesara, J. et al. Integrating machine learning and blockchain to develop a system to veto the forgeries and provide efficient results in the education sector. Vis. Comput. Ind. Biomed (D.). Art 4, 18 (2021). <https://doi.org/10.1186/s42492-021-00084-y>
5. B. K. Sethi, P. K. Sarangi and A. S. Aashrith, "Medical Insurance Fraud Detection Based on BlockChain and Machine Learning Approach," 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 2022, pp. 1-4, doi: 10.1109/ICERECT56837.2022.10060811.
6. Goyal, Adit & Elhence, Anubhav & Chamola, Vinay & Sikdar, Biplab. (2021). A Blockchain and Machine Learning based Framework for Efficient Health Insurance Management. 511-515. 10.1145/3485730.3493685.
7. Abbas, K., Afaq, M., Ahmed Khan, T., & Song, W. (2020). A Blockchain and Machine Learning-Based Drug Supply Chain Management and Recommendation System for Smart Pharmaceutical Industry. Electronics. <https://doi.org/10.3390/electronics9050852>
8. Y. Liu, F. R. Yu, X. Li, H. Ji and V. C. M. Leung, "Blockchain and Machine Learning for Communications and Networking Systems," in IEEE Communications Surveys & Tutorials, vol. 22, no. 2, pp. 1392-1431, Second quarter 2020, doi: 10.1109/COMST.2020.2975911.
9. Qiu Z, Xie Z, Jiang X, Ran C, Chen K. Novel Blockchain and Zero-Knowledge Proof Technology-Driven Car Insurance. Electronics. 2023; 12(18):3869. <https://doi.org/10.3390/electronics12183869>
10. S. Jain, A. Anand, A. Gupta, K. Awasthi, S. Gujrati and J. Channegowda, "Blockchain and Machine Learning in Health Care and Management," 2020 International Conference on Mainstreaming BlockChain Implementation (ICOMBI), Bengaluru, India, 2020, pp. 1-5, doi: 10.23919/ICOMBI48604.2020.9203483.
11. Ashfaq T, Khalid R, Yahaya AS, Aslam S, Azar AT, Alsafari S, Hameed IA. A Machine Learning and Blockchain Based Efficient Fraud Detection Mechanism. *Sensors*. 2022; 22(19):7162. https://doi.org/10.3390/s22197162
12. Kayikci, S., Khoshgoftaar, T.M. Blockchain meets machine learning: a survey. J Big Data 11, 9 (2024). <https://doi.org/10.1186/s40537-023-00852-y>
13. Agrawal, D., Bansal, R., Fernandez, T.F., Tyagi, A.K. (2022). Blockchain Integrated Machine Learning for Training Autonomous Cars. In: Abraham, A., et al. Hybrid Intelligent Systems. HIS 2021. Lecture Notes in Networks and Systems, vol 420. Springer, Cham. <https://doi.org/10.1007/978-3-030-96305-7_4>
14. .Nouhaila El AkramI, Mohamed Hanine, Emmanuel Soriano Flores, Daniel Gavilanes Aray, and Imran Ashraf “ Unleashing the Potential of Blockchain and Machine Learning: Insights and EmergingTrends From Bibliometric Analysis “ <https://ieeexplore.ieee.org/document/10192385/authors#authors>
15. .Sudeep Tanwar , Qasim Bhatia , PruthvI Patel, Aparna Kumari, Pradeep Kumar Singh and Wei-Chiang Hong” Machine Learning Adoption in Blockchain-Based Smart Applications: The Challenges, and a Way Forward“ <https://ieeexplore.ieee.org/document/8938741/authors#full-text-header>